On the (In)fidelity and Sensitivity of Explanations

Chih-Kuan Yeh, Cheng-Yu Hsieh,
Arun Sai Suggala, David I. Inouye, Pradeep Ravikumar

Objective Measure

Infidelity — expected difference between the function value change after perturbation $f(x) - f(x - I)$ and explanation dot perturbation $I \cdot \Phi(f, x)$

— Many explanations optimizes infidelity with respect to some perturbations
— We may calculate the closed form optimal solution
— We can design new explanations by defining new perturbations

perturbation	distance to baseline	ϵ · coordinate basis vector	$\mathbb{P}(Z=z) \propto \frac{d-1}{\binom{d}{\|z\|_1}\|z\|_1(d-\|z\|_1)}$	distance to (baseline + Gaussian noise)	square block in image
explanation	IG [1] LRP [2] DeepLift [3]	Gradient [4]	Shapley value [5]	NB	Square

Sensitivity — expected function value change after small perturbation in input

— If explanation is "sensitive", it may undermine the credibility of explanations [6]
— We show that we can improve both the sensitivity and infidelity of a given explanation by smoothing explanations

Theorem 4.1. *Given a black-box function* $\mathbf{f}$*, explanation functional* Φ*, the smoothed explanation functional* Φ_k*,*

$$\text{SENS}_{\text{MAX}}(\Phi_k, \mathbf{f}, \mathbf{x}, r) \leqslant \int_{\mathbf{z}} \text{SENS}_{\text{MAX}}(\Phi, \mathbf{f}, \mathbf{z}, r) k(\mathbf{x}, \mathbf{z}) d\mathbf{z}.$$

Thus, when the sensitivity SENS_{MAX} is large only along some directions $\mathbf{z}$, the averaged sensitivity could be much smaller than the worst case sensitivity over directions $\mathbf{z}$.

Theorem 4.2. *Given a black-box function* $\mathbf{f}$*, explanation functional* Φ*, the smoothed explanation functional* Φ_k*, some perturbation of interest* $\mathbf{I}$*,* C_1, C_2 *defined in (6) and (7) where* $C_1 \leqslant \frac{1}{\sqrt{2}}$*,*

$$\mathit{INFD}(\Phi_k, \mathbf{f}, \mathbf{x}) \leqslant \frac{C_2}{1 - 2\sqrt{C_1}} \int_{\mathbf{z}} \mathit{INFD}(\Phi, \mathbf{f}, \mathbf{z}) k(\mathbf{x}, \mathbf{z}) d\mathbf{z}.$$

We propose objective evaluation metrics to quantify the infidelity and sensitivity for feature-based explanations. We show that we can improve the infidelity and sensitivity simultaneously for a given explanation by smoothing.

Carnegie Mellon University
MACHINE LEARNING DEPARTMENT
PURDUE UNIVERSITY
College of Engineering

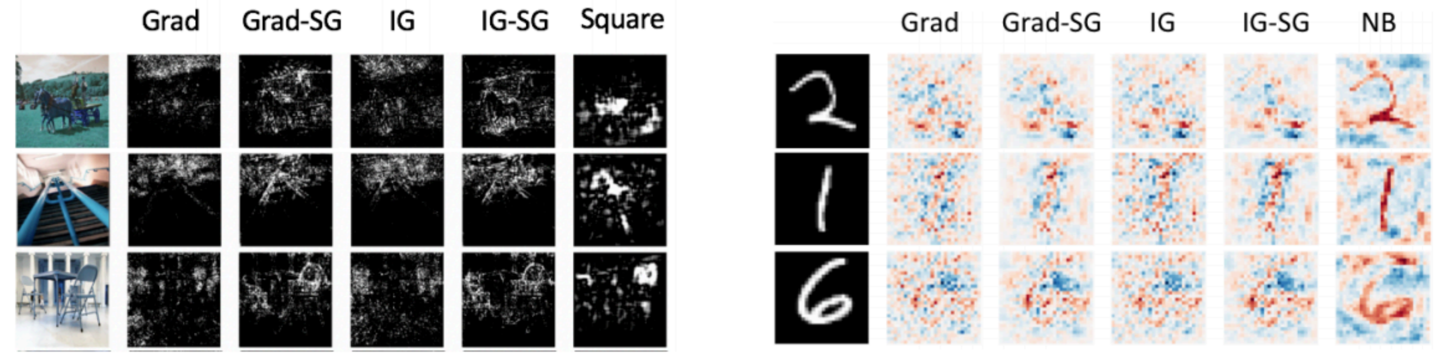
Experiments

Infidelity and Sensitivity

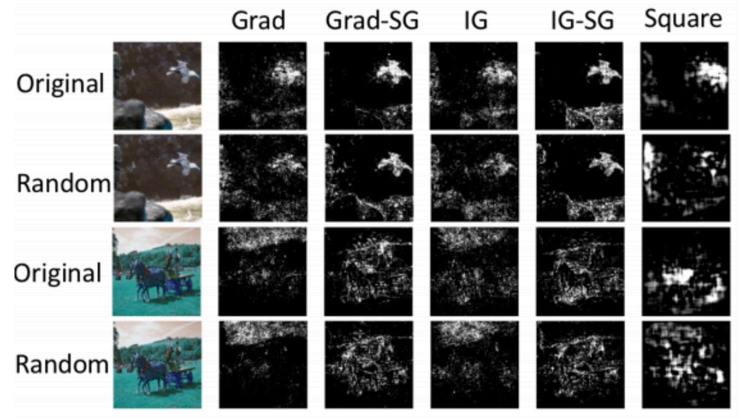
Datasets	MNIST	
Methods	SENS_{MAX}	INFD
Grad	0.86	4.12
Grad-SG	0.23	1.84
IG	0.77	2.75
IG-SG	0.22	1.52
GBP	0.85	4.13
GBP-SG	0.23	1.84
Noisy Baseline	0.35	0.51

Datasets	MNIST		Cifar-10		Imagenet	
Methods	SENS_{MAX}	INFD	SENS_{MAX}	INFD	SENS_{MAX}	INFD
Grad	0.56	2.38	1.15	15.99	1.16	0.25
Grad-SG	0.28	1.89	1.15	13.94	0.59	0.24
IG	0.47	1.88	1.08	16.03	0.93	0.24
IG-SG	0.26	1.72	0.90	15.90	0.48	0.23
GBP	0.58	2.38	1.18	15.99	1.09	0.15
GBP-SG	0.29	1.88	1.15	13.93	0.41	0.15
SHAP	0.35	1.20	0.93	5.78	–	–
Square	0.24	0.46	0.99	2.27	1.33	0.04

Visual Example



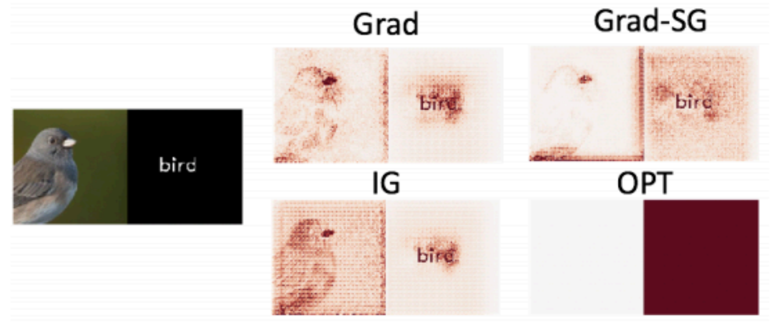
Sanity Check



	Grad	Grad-SG	IG	IG-SG	Square
Corr	0.17	0.10	0.18	0.16	0.13
Corr (abs)	0.57	0.62	0.61	0.62	0.28

Table 2: Correlation of the explanation between the original model randomized model for the sanity check.

Human Evaluation



	Grad	Grad-SG	IG	OPT
Infid.	0.55	0.38	0.35	0.00
Acc.	0.47	0.50	0.53	0.88

Table 3: The infidelity and the accuracy human are able to predict the input blocked used based on the explanations.

Reference

[1] Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In ICML 2017.
[2] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7):e0130140, 2015.
[3] Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In ICML 2017.
[4] Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. Not just a black box: Learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713, 2016.
[5] Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In NeurIPS 2017.
[6] Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile. In AAAI 2019.